

ASSADIAN et al
U.S. National Phase of PCT/GB2004/004028

AMENDMENTS TO THE ABSTRACT

Please insert the Abstract of the Disclosure which is on the attached sheet.

ABSTRACT OF THE DISCLOSURE

A method and apparatus are provided for generating, from an input set of documents, a word replaceability matrix defining semantic similarity between words occurring in the input document set. For each word, distinct word sequences of predetermined length are identified from the documents of the set, each word sequence being indicative of the context in which the word was used and, according to the relative frequency of occurrence of the identified word sequences for the word, fuzzy sets are generated for each word comprising membership values for corresponding groups of word sequences. For each pair of words occurring in the document set, their respective fuzzy sets are used to calculate the probability that the first word of a pair is semantically suitable as a replacement for the second word of the pair, these probabilities being collated to form a word similarity matrix for use in an improved method of determining document similarity and in information retrieval.